doi: 10.36689/uhk/hed/2019-02-017

Open Science, Open Research Data and some Open Questions

Jakub NOVOTNÝ

University of Economics, Prague, Czech Republic
novotnyj@vse.cz

Abstract. Open access to research data is one of the key themes of current science and relevant research and development strategies at least in Europe. A systemic change in the modus operandi of science and research should lead to so-called Open Science. This overview paper presents the principles of Open Science and in detail "FAIR data" as one of the key assumptions for open-access research data. Research data meeting the criteria FAIR data must be findable, accessible, interoperable and re-usable especially for machines. Data elements are viewed in modular way - data elements, persistent identifier and metadata. There are, as a final discussion, mentioned three areas crucial for further elaboration and application of concept of open research data: elaboration of sophisticated methodology; analysis of financial aspects of open research data for research institutions; and linkage with the assessment of research institutions as an important incentive.

Keywords: Open Science, Open Research Data, FAIR Data.

1 Introduction

Science and its methodology develops over time. In the history of modern science, this process was primarily an internal affair of science and the development of its paradigms. In recent decades, however, the methodology of science has more than ever been significantly influenced by technological developments, which nevertheless relate to scientific knowledge, and to the science policy that emerges from general politics.

Science and research activities rely by many ways on ICT and in same time research generate inexhaustible quantity of digital content. This digitalization brings new challenges and one of nowadays paradigms is openness - open access to research publications and research data. This paradigm is called Open Science. The production of research data is growing significantly every year, bringing a number of challenges and challenges not only in terms of their processing and accessibility but also in the field of science methodology itself. Efficient processing of research data is possible only by machine technology, which brings additional demands on digital scientific data.

In this overview paper, we will focus on the concept of FAIR data, a set of basic (minimum) policies and practices designed by the eScience community that allows people and machines to easily search, access, collaborate, and reuse research data. The aim of the paper is to approach the individual principles of FAIR data within Open Science, the modular concept of a data object based on these principles and, in particular, to point out some possibly problematic issues of open research data and the FAIR principle .

2 Open Science

Science and its individual disciplines are evolving. There is also a change in the way of its operation and the source and mechanism of science funding. Technological development (which in itself is the result of scientific knowledge) offers new tools for research and dissemination and publication of results. Digital technologies offer a faster and cheaper way of presenting results than before. The society's relationship to science and the development of science is also changing. Science and its strategic development have become part of wider political concepts.

In the last decade we can therefore meet the concept of Open Science (or earlier Science 2.0) with three main attributes [4]:

- A significant increase of scientific production, open research and remote collaboration and online (open) access to scientific information.
- An emergence of data-intensive science by availability of large-scale datasets (petabytes) and by high performance computing.
- An increase in the number of actors in science.

Open Science is therefore a systematic change in the modus of operandi of research activities and is affecting the research cycle and all of its stakeholders. Research process in open form is shown in the following figure.

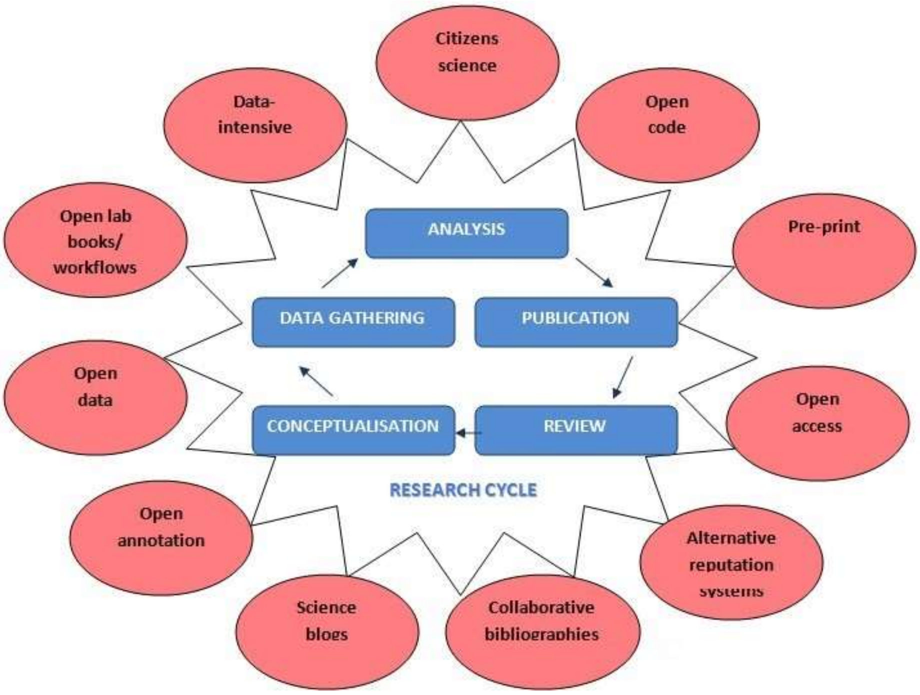

Fig. 1. Open Science trends [4].

Although there are many Open Science trends, open access is the most discussed and elaborate one, at least in terms of science policy within the European Union. According to [3] "open access refers to the practice of providing online access to scientific information that is free of charge to the end-user and reusable" and scientific information are divided to categories: peer-reviewed scientific research articles and research data. It is very important how access is defined. Access is not only "the right to read, download and print – but also the right to copy, distribute, search, link, crawl and mine."

Research data in this context are data (statistics, results of experiments, measurements, observations, survey results, interview recordings and images) in digital form allowing users to freely access, mine, exploit, reproduce and disseminate them. Open access to scientific publications and research data will according to European science policy (especially through Horizon 2020) improved quality of results, encourage collaboration, avoid duplication of effort, speed up innovation and involve citizens and society to science.

3 FAIR Data

A much more detailed specification of research data features within the Open Data concept is the so-called FAIR Data principle. The basic document dealing with FAIR Data is the Guidelines on FAIR Data Management, which specifically addresses the recommendations for the Horizon 2020 R & D beneficiary or the participants involved in the Open Research Data Pilot, but its impact on the scientific community is wider and touches the issue of openly accessible scientific data in general. The guide does not detail the principle of FAIR data. It contains only an initial indication that it helps the beneficiaries to make their research data findable, accessible, interoperable and reusable (FAIR) and also states in the annex that the research data should comply with the FAIR principles, and refers to FORCE11 and a published article in Nature [7] for further details. The FAIR Data principle is built on work of the Concept Web Alliance and the Joint Declaration of Data Citation Principles with no direct reference to any theoretical concept of metadata.

So let's look at the FAIR data concept. The principles are not only related to the data itself (in a strict definition), but also to the research procedures, algorithms and tools that lead to the production of such data. In the basic breakdown, there are 15 principles or recommendations that research data should meet:

To be Findable:

F1. (meta)data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata (defined by R1 below)

F3. metadata clearly and explicitly include the identifier of the data it describes

F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

To be Reusable:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with detailed provenance

R1.3. (meta)data meet domain-relevant community standards

These policies should not serve as a standard or specification, nor does it address the technical implementation of the data produced and stored. Their intention is to assist scientific institutions and research teams in deciding on a specific way of realizing the digital outputs of their research so that these outputs can be searched, accessed, involved in further research, and further exploited within the scientific community (and not only). It is in fact an explanation of the scientific-methodological requirements for digital outputs of scientific work and the scientific and methodological assumption of machine evaluation and mining of research data.

The core of FAIR data is the internal structure of data and data objects in terms of subsequent practical steps with regard to the formatting and publication of research data. Given that the main objective of FAIR data is that data is findable, accessible, interoperable and reusable (FAIR for machines), data is limited to digital data only.

Each data object is therefore a digital object where the data object is defined as an identifiable data item consisting of data elements, metadata and identifier. The smallest data object is a simple identifier that refers to a concept (ie, an idea or a "unit of thought") that does not have the nature of a digital object. Each data object should then, according to FAIR principles, include at least one persistent identifier (PID).

The FAIR principle requirement is that each data object as a whole is assigned a PID and at least a minimum set of metadata about the given data object. The data object may contain its own intrinsic and user defined metadata and contain from one to a large number of data elements. The modularity of the data object is shown by the following figure.

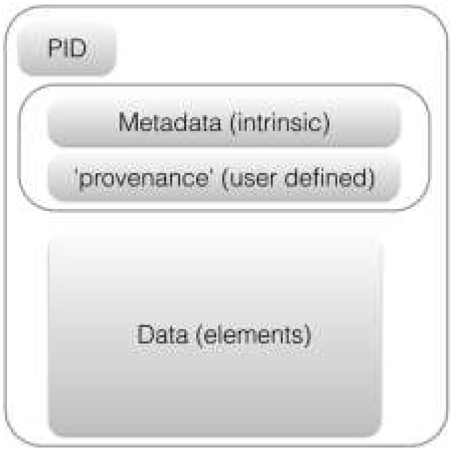

Fig. 2. Modularity of a data object [6].

The individual data element can then be used, quoted or distributed separately as a new data object with a new PID and correspondingly selecting the metadata of the original data object so that traceable binding and citation to the original data object is traceable.

Storage of scientific data for machine-readable data (so-called FAIRports) must meet the following conditions:

- Include FAIR data objects (verified by appropriate authority)
- Provide given data objects under strictly defined accessibility conditions for reuse
- Provide a complete and open-access description of all technologies, dictionaries, and data formats used

FAIRport must also contain data objects at least level 1 in terms of their FAIRness (FAIRness). Explicitly level 1-4 is described, but from the next description we can extend this range by two transitions to:

- Level -1 = There are data objects in the repository that do not have PIDs or their own internal metadata
- Level 0 = Each data object has a PID, but some data objects do not have their own internal metadata
- Level 1 = Each data object has both PID and its own internal metadata
- Level 2 = Each data object has a full FAIR annotation, ie its own internal metadata and user defined metadata demonstrating the origin of the data elements in the FAIR format
- Level 3 = Data elements in data objects technically meet FAIR principles but are not fully open and reusable
- Level 4 = All data elements and metadata meet the FAIR policy conditions and are completely publicly accessible for a clear licensing.

Graphics of these 6 levels are captured by the following picture:

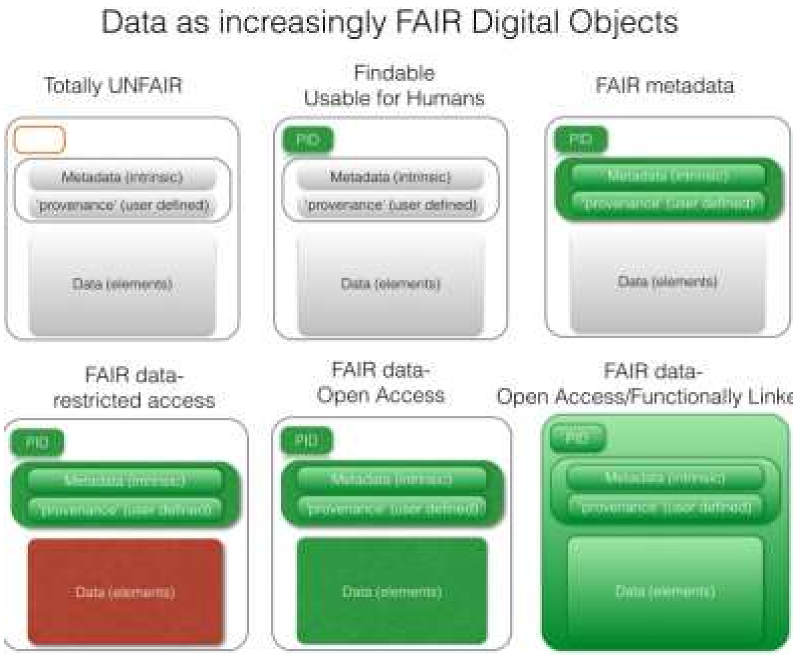

Fig. 3. Levels of FAIRness [6].

4 Conclusion - Open Questions for Open Data

In the presented text we have been devoted to the characterization of the concept of Open Science and in greater detail the principle of the FAIR data and its internal modularity. As a final discussion, it is advisable to mention some of areas that we consider to be crucial for further elaboration of these concepts and, in particular, their introduction into scientific practice.

Documents at both European and national levels contain only a general postulation of the Open Data principles, or refer to other working documents on the topic. Therefore, a sophisticated methodology is missing. In this respect, research community outputs (eg via the Research Data Alliance) rather than official bodies can be expected. In case of the Czech Republic, Association of Libraries of Czech Universities (ALCU) signed in year 2012 Berlin declaration and publish ALCU Open Access Policy. Subsequently, on June 14, 2017, the Czech National Strategy for Open Access to Research Information for 2017-2020 has been approved by the Government of the Czech Republic. So the Open Science principles are there gradually implemented at a very general political level.

Although the obligation of open research data (which complies with FAIR data principles) is already pilot-tested and is becoming a requirement of many science-funded programs, there are no studies to analyze both direct and indirect financial aspects of open research data for research institutions and benefits).

Promoting open research data must be linked to the assessment of research institutions and research teams as a major incentive. Moreover, if it is properly matched with funding, it will be an effort of most research teams to provide their research data in an open standardized manner.

In spite of the open questions linked to open research data, it is certain that this is one of the major trends in science policy and science will become more open than in the past and the paradigm of open science, including the principles of machine processing of scientific data, will transform it and transform its methodology.

Acknowledgements. This article was created with the subsidy of the means of institutional support for long-term conceptual development of science and research at Faculty of Informatics and Statistics, University of Economics, Prague.

References

1. European Commission: Access to and Preservation of Scientific Information in Europe. Report on the implementation of Commission Recommendation. EC 2012. 4890 final.
2. European Commission: Guidelines on FAIR Data Management in Horizon 2020. EC 2016. Version 3.0.
3. European Commission: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. EC 2017. Version 3.2.

4. European Commission: Public Consultation ‘Science 2.0’: Science in Transition. EC 2014, from http://ec.europa.eu/research/consultations/science-2.0/background.pdf, last accessed 2018/10/12.
5. Fabián, O. (2013) "Open access in the Czech Republic: an overview", Library Review, Vol. 62 Issue: 4/5, pp. 211-223, https://doi.org/10.1108/LR-09-2012-0096
6. FORCE11: Guiding principles for findable, accessible, interoperable and re-usable data, publishing version B1.0. from https://www.force11.org/fairprinciples, last accessed 2018/09/29.
7. Wilkinson, M. D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18